12

Chinese Full-Text Retrieval Technology Applied in National Occupational Training and Testing Network in WWW Environment

Meng Qingyuan
department of computer & information, Beijing Communications Management Institute for Executives
Beijing East Yanjiao, Zip: 101601, Tel: (010)64625507-8399, E-mail: mengqqii@public.bta.net.cn
Chen Min
department of computer & information, Beijing Communications Management Institute for Executives
Beijing East Yanjiao, Zip: 101601, Tel: (010)64625507-8399, E-mail: startzx@public.bta.net.cn
Li Jingsheng
Occupational Skill Testing Authority, Ministry Of Labor P. R. China
17 Huixing, XiJie Beijing P. R. China, Zip: 100029, Tel: (010)64963244, Fax: 64915804, E-mail: ostaml@netchina.com.cn
Zhang Linzhi
Occupational Skill Testing Authority, Ministry Of Labor P. R. China
17 Huixing, XiJie Beijing P. R. China, Zip: 100029, Tel: (010)64963244, Fax: 64915804, E-mail: ostaml@netchina.com.cn
Dong Shuming
Beijing Nation Computer Company
54 Baishijiao, Haidian, Beijing, Zip: 100044, Tel: (010)68312249, Fax: 68312249, E-mail: nation@hn.ciiss.net.cn

Information Systems in the WWW Environment C. Rolland, Y. Chen & M. Fang (Eds.)
© 1998 IFIP. Published by Chapman & Hall

Abstract

This paper describes how CGI and API techniques are used to realize fast indexing and remote retrieving vast quantity original data on personal computers.

In order to make an overall management and high-speed holographic research on the data about national occupational skill training and testing, we've made this new rough describing textual data structure by combining with intelligent word-splitting technology and word-approaching fuzzy researching calculation method to have a full-text intellectual multimedia research on the non-structural holo-data; to have a full-text intellectual multimedia research on structural holo-data by adopting intellectual double-matching quadratic indexing skill; to use the common PC to create an united intellectual concurrent index level----a chaos indexing structure to have a high speed intellectual multimedia full-textual research under the environment WWW and to make simultaneous high speed searching for versions information, super version information and database information by using the synthetic application technologies of CGI and API according to the index structure which is set up according to the Chaos theory and Fractal technology .

Keywords

Full-text Retrieval indexing Multimedia Chaos

1. SYSTEM DESCRIPTION

National Occupational Training and Testing Network (NOTTnet) is a fully functional network system specialized in career training, certification, and development. It has been jointly developed and implemented by Department of Occupational Skill Development and Occupational Skill Testing Authority (OSTA), Ministry of Labor. The purpose of developing such a system is to improve career training programs in China.

NOTTnet includes the following sub-systems and components:

- Dynamic information dissemination system
- Occupation certification information management system
- Training and certification expert system
- Training institution management information system
- Training materials information system
- Human resources and experts information system
- Occupation standards/codes and laws/regulations system
- Communication system
- Other expandable sub-systems

The above systems enable NOTTnet to provide to the users in government branches and all paths of the society with the functions of dynamically disseminate information on, and accept, process, and manage information queries in careers, certification, experts, professionals, training programs, training facilities, training materials, and training and certification standards. It also manages and provides services concerning test materials, test results and test statistics.

NOTTnet operates and is managed and maintained at three levels: national level; province level; and city/region level. The Occupational Skill Testing Authority (OSTA) , Ministry of Labor is responsible for the NOTTnet's overall operations and to manage and maintain its central databases. At the second and third level, each local centers are responsible for local operation, including data collection, user services, and database maintenance.

2. TECHNICAL FEATURES

NOTTnet is a comprehensive data management system. It is end-user oriented and its data are shared by a large number of users throughout the entire country. Its vast databases cover many subject domains and are managed at different levels.

The fundamental difficulty in developing and implementing a system in such a large scale is the technique of rapid indexing on large quantity data and fast retrieval of such data in a WWW remote mode.

NOTTnet must have the following features:

(1) Data source inheritance and compatibility. The system must be able to use local levels' original data and summarized data and reports based on such summarized data from city, region, and province levels. It must be able to retrieve

and directly use these data files.
(2) Effective management of large quantity data. The system's data deal with a labor force of 100 million employees in more than 5,000 labor categories.
(3) Full range data processing. The system's data include structureless text data and structural digital data. These data are processed through comprehensive multimedia methods.
(4) Fast processing speed in WWW environment. Fast speed processing is required in order to enable the data sharing throughout the entire country.

In order to overcome the above mentioned fundamental difficulty in the system development, NOTTnet must adopt the following techniques:

(1) Full-text indexing of English/Chinese mixed text data;
(2) Relevant multimedia date connecting;
(3) Network file server sharing;
(4) Remote client/server;
(5) Mixed data processing in WWW platform, including hypertext data format and structural data format.

NOTTnet uses a unique outline description structure to process data in natural language. Individual characters as the basic data units are organized into indexed tables. An intelligent phrase filtering technique and a fuzzy indexing algorithm are also used in data processing to deal with the phrases with close meanings. Different types of text data can be directly input into database without any manual or mechanic identification. Data stored in database are available for fast retrieval with the same indexing logic and by both Chinese and English keys. Multimedia data such as sound files, graphics, video clips, etc. are comprehensively handled by the system. It also introduced an intelligence based technique of secondary full-text indexing on structured data. According to the chaos theory and Fractal principle and using combined CGI and API technologies, the system has realized comprehensive and fast indexing on, and remote retrieval of, extremely large datasets.

3. TECHNICAL IMPLEMENTATION OF THE PROJECT

Outline structure indexing and fuzzy approaching retrieval are two new text data description. Although these two are independent to each other, they are mutually supportive. The key feature of these two structures is that they do not tend to accurately specify or describe data, but merely regulate a direction for further specification or description. Within the range of direction and in response to the contexts, the computers are utilizing their intelligence and dynamically approaching the accuracy. The intelligent secondary indexing technique processes structured data to establish intelligent matching indexes, and creates outline structures for these indices. Full-text indexing on structured data is realized through the

application of these two techniques. Chaos indexing structure adds indexing parameters to the outline structure. At the query level, chaotic query requests are arranged into certain orders. This enables the system to realize remote, full range indexing on extremely large amount of original data.

A series of issue must be addressed in order to realize the rapid and remote indexing and retrieving large amount raw data on the World Wide Web. First of all, the system needs to deal with indexing and retrieving non-structured text data. Secondly, it needs to link multimedia data and their related text data together. Then it needs to utilize the intelligent matching secondary indexing mechanism to index and retrieve structured data. The last but not the least, it needs to do joint indexing and hypertext data indexing, with the help of CGI and API techniques. NOTTnet system emphasizes the full-text indexing method as its primary key technique. As the bottom layer, it sets up the raw data as its foundation databases. At the top layer, the system takes the users querying interface as its highest operation requirements. Between the bottom data layer and the top users query layer, the system utilizes the above mentioned processing techniques and their variations to channel the interactive communications of the top and the bottom layers. The system objective is thus achieved.

(1) Processing non-structured text data with the outline description data structure.

In NOTTnet system, we did not take as granted various popular data processing techniques. Popular database techniques are merely capable to process structured data, and are having problems such as slower speed, lower comprehensive rate and accuracy, and are less actually used. With these problems, they can hardly process raw data and is not suitable for NOTTnet system.

The new, intelligent full-text indexing technique handles both Chinese and English text data and multimedia data as well. Raw data can be used directly and retrieved randomly and instantly. With the technique the system can reach full range data processing of raw data.

The full-text indexing technique has solved primarily the problem of fast indexing on and retrieval from non-structured data. The concept of full-text indexing is structure-insensitive, any character or word included in a document can be queried and retrieved rapidly, without distinguishing their structure. The most difficult part of such an indexing technique is that its speed must be very fast, its coefficient of expansion very small, its error rate very low, and its accuracy very high. The key issue here is how to establish the indexing/retrieving speed. After our careful analyses and research and large quantity of calculation and tests, we achieved a scheme which is a combination of individual character indexing and intelligent word/phrase indexing. This scheme focuses primarily on individual characters as the basic data unit, and is enriched by the intelligence based word/phrase distinguishing technique to fulfill full-text indexing/retrieving requirements. This scheme has created two core techniques, the first one is a new fast indexing

algorithm: an algorithm through a fuzzy search to approach toward a target word; the second, a new text data indexing structure named outline structure.

Natural raw data do not possess any structure. If index and retrieval are applied to these data, certain structured, indexed "meta-data" documents must be created to describe such data. Due to the features of Chinese language, either a character or a word or phrase can be chosen as the basic index unit on which the index system is established. Since Chinese language actually uses individual characters as its smallest linguistic unit, if word or phrase is elected to as the basic indexing unit, the indices established may encounter the problems such as hard to distinguish words with close meanings, easy to miss query target, massive work is required to maintain the database, large resources are demanded to store data, etc. Given the above, in our project, we selected individual character as the basic indexing unit, and decided using distributive table technique to construct data'soutline structure. This plan has proved the following benefits:

more flexibility;

· technically easier to implement;

· lower development cost;

· convenient to use; and

· higher accuracy and lower error rate, reached with the help of the dynamic word distinguishing technique.

The outline structure is another new hightech. Though it can be separated from the fuzzy approaching algorithm, the outline structure technique is actually used hand in hand with it. The characteristic of the outline structure is: it does not intend to specify raw data accurately, nor does it actually describe the data in detail. Rather, it roughly describes the data and merely indicates the order and frequency of individual characters contained in a data set. In the process of retrieval, computers utilize their intelligence to automatically approximate and dynamically approach towards the accurate results, according to the query contexts, query conditions, and the logical relationships among all the conditions involved. Although fuzzy, this searching process leads to a higher rate of accuracy. As index ration is 1:0.33, the data search speed reaches 0.4 second per million individual characters. When the indexing ration is raised to 1:1.33, the speed is increased to 0.35 seconds per 10 million characters. In terms of indexing speed, it takes less than 20 minutes to index 10 million individual characters, when dynamic random data updates are also allowed.

(2)Establishing links between multimedia data and text data with an automatic linking method while populating the database.

Multimedia raw date exist in the format of graphics, sounds, and movie clips, which are independent of each other. They're contents cannot be indexed or retrieved at high speeds. These data, however, are often times the attachment to certain text data. This fact enables an indexing on two groups of data's relevancy; thus, when the text data and their relevant multimedia data are collected and stored into databases, the multimedia data's file information is added to the text data's index files. This process automatically establishes linkages between text data and multimedia data, based on their relevancy. In the retrieval of text data files, the relevant multimedia files are also accesses and their contents displayed or played.

(3) Storaging, indexing, and retrieving vast datasets through the hierarchy classification (data tree).

When raw data are collected and moved into databases, multiple level hierarchical data classification is applied and a data tree is formed, according to data's characteristics.

Cross-references and chain links and constructed onto the classified data. As a result, the data is stored distributively on the data tree, but retrieved and accessed through references and links.

(4) Full-text indexing on structured data with intelligent matching secondary indexing.

In order to accurately and quickly use structured data, we did not use the direct indexing on, and direct retrieval from, databases. We instead elected to create a fast indexed shadow document to reflect the data stored in a database. The indexed data in the shadow document becomes structured, and are further rapidly retrieved through the full-text indexing technique. This data set serves as an intermediate data set to generate the outline structure of this intelligent matching index document. When the full-text data are retrieved, still the fuzzy searching algorithm and the outline structure are used to approach towards the final, accurate results. In the entire process, database document consists of raw data, while the full-text retrieval actually accesses another set of data, i.e., the intelligent matching index data. .DBF files are not touched and modified in any sense, only the index data document is queried and searched. When the query results are to be displayed, the index data document is coordinated with the paragraph content description in order to retrieve the contents stored in .DBF documents. A full-text indexing and retrieving is thus realized.

Characteristics of the WWW data retrieval

The scheme of the WWW is the combination of browser and server. It adopts HTTP protocol and uses HTML language to generate files. This scheme is an extended and distributive client/server structure. This structure enables all WWW users to share common multimedia browsers and realizes cross platform data processing. Between a client (browser) and WWW server exists the HTTP protocol. This protocol establishes itself from a no-connection condition, and is thus a suitable protocol for multiple, simultaneous queries and retrievals. When established upon a query request, a protocol moves out of the no-connection state with an instant connection. As soon as it has processed the first request, it is immediately released and returns to a ready status for processing the second request. In this process, the server side is almost always in a ready state, or statusless state . A crucial point is that the protocol does not have its own memory. If a series of queries are sent, and the latter query handling depends on the former query handling, the intermediate data must be stored outside of the protocol. In order to satisfy each and every query request, the server should be able to process 20 or more requests per second, no matter the same or different data sets are accessed.

HTTP allows interoperability between the client side and the server side. As data are being transferred, a selectable meta-data prefix is added to the primary data. This enables the client machine to send the meta-data to the server along with the query results. Consequently, a query request becomes a conditional request, and the server is able to provide information on what is being transferred. Its service to the clients is thus enhanced.

In the WWW environment, files used are in HTML format. HTML is the specialized language for construction of multimedia files. HTML files are plain text

files or full ASCII files, which have a HTML or HTM file extension. One of its unique features is that at the beginning and the end of a file, HTML markup symbols are used. With its marks and properties, HTML describes the semantic of a text paragraph. In addition, it provides hypertext links between two files or two parts of a file. The hypertext links have made surfing the WWW possible.

If the data, which are being retrieved in a browser, are not in the HTML format but in some other format such as .DBF, some extended application programs on the server must play the role of "midway house." They first convert the data into HTML files and then send these files to the user's browser.

The most popular "midway house" techniques include CGI and API. CGI and API realize the interoperations of the client side and the server side. The server side can send information to CGI programs, and CGI can feed information back to the server. CGI allows a WWW server to run external application programs. A server capable of running external programs can utilize external resources such as external databases, and then generate HTML files, which are sent back to a user's browser. API, on the other hand, allows developers to write server plug-ins and extensions to enhance the service functions. In comparison to CGI, API application programs are better integrated into the server's functions. API consumes little system resources so that its higher operation efficiency further enhances the system performance and system security.

(5) Simultaneous query handling in WWW environment with the chaos structure

NOTTnet system's raw data are text documents or database documents, most of which are attached with multimedia data. In WWW environment, rapid indexing and retrieving post three requirements:

· automatic formation of hypertext;
· fast index and retrieval speed; and
· single visit query handling.

Common Gateway Interface (CGI) and Application Programming Interface (API) techniques are used as an intermediate layer between the bottom data layer and the top user query layer. The CGI/API layer constructs a chaos index structure, a unified intelligent joint indexing/retrieving tier, which enables the system to respond to high speed simultaneous queries tasks. Through this step, intelligent multimedia full-text indexing and retrieving techniques, which used to be possible only on PCs, LANs, or in remote accessing conditions, are now upgraded also working for WWW environment.

In WWW environment, full-text indexing and retrieving are required to comply with the special situations such as no-linkage, statusless, concurrent retrieval, and high speed simultaneous query tasks, etc. This implies that each query task is a one-time visit event and can only be handled as a single visit event. This also means no intermediate conditions are saved, nor initialization operation is carried out, and an instant "first in and first out" pattern is followed all the time. The reason for this is

obvious: the next query visit is likely a query from another user. Consequently, each query request is must carry its complete query instructions; and the query result from the server is supposed to have a complete query result description.

In this manner, the system is able to fulfill the requirements posted to it by a continuous query flow, and the system is able to be constantly prepared for the next query visit.

To fulfill these requirements, the system needs a special design of an index data paragraph, which matches the complete query instruction. This kind of matched structure organizes the query requests into a certain sequence for processing, although they look completely irrelevant and in chaos. In our terminology, we call this kind of matching structure the structure of chaos. It is based on the theory of chaos and fractal principle. On the basis of the outline structure, a few descriptive paragraphs are added to, and its query requirement and forwarding position are included in these paragraphs.

NOTTnet is a native 32-bit system. Its core programs have been developed and are running in the Windows NT environment. When implemented on the WWW, a 32-bit central database has been placed on the WWW server. Its retrieval is through the Internet browsers such as MS Internet Explorer and Netscape. This is consistent with the client side. A CGI program is also used for reinforcing the WWW server standards and full-text indexing and retrieving regulated by these standards. The NOTTnet system implemented in this manner has obtained a complete framework, which fully complies with the Internet criteria.

1) The chaos structure enables the rapid query and retrieval in the WWW environment

The chaos structure is the breakthrough technique in the development of the NOTTnet system in the WWW environment. The chaos structure is constructed on the system's outline structure as its foundation. Some descriptive paragraphs are added to the chaos structure to form a indexing/retrieving layer. When interactive query requirements are forwarded to the system through the Internet browser, the CGI programs accurately allocate the primary keys on the indexing/retrieving layer, and these keys further lead to the raw data stored in the data documents at the foundation layer. It is at the indexing/retrieving layer that the query results are determined (target document) and the results are processed and transferred back to the querying browser for user's display.

The actual process is as follows:

When raw data are collected into database, the outline structured text data documents are created from all types of documents, including hypertext documents, plain text documents, and database documents. The outline structured documents are for full text indexing and retrieval.

Each user input dataset is mounted to a descriptive paragraph before the datasets are stored onto the WWW server machine. This process converts the outline structured text data into chaos structured index documents. The API modules,

located at the full text retrievable database foundation, are also loaded into the WWW server, which are responsible for managing various types of raw data documents in the databases.

When data retrieving takes place, a user gives query instruction data through an I/O terminal. The querying data travel through certain network connection to reach the WWW server. The server side watchdog program is triggered by the CGI request, and the system starts to process the querying instructions. According to the chaos structured information format, the received querying data are translated into querying parameters to be attached to the CGI programs. With the querying parameters, CGI programs can instantly allocate and identify the query result from the chaos structure documents, and can further forward to API programs the intermediate query results and querying parameters. The API programs are running in the background and are further processing these querying data at the server's foundation database level.

At this point, the CGI programs have finished their querying task, they quit form this task and let the CGI programs take over. Comprehending the intermediate query results and attached parameters, the API programs conduct the separating and decoding routines over the raw, outline structured data. Target data are selected at the end of the processing routines. All documents which are not in hypertext format are automatically converted to the HTML format. According to the querying parameters carried over, the API programs bring up and send the query responses back to the querying user terminal. This marks a query process concludes.

2) The chaos structure provides solutions to the CGI programs' low efficiency problem.

Although CGI programs are powerful in interactive query functions, the CGI programs and the WWW server take moves differently. Generally speaking, each CGI program can only process one query task. Hence each query will lead to a new CGI instance to serve the process. When the amount of query tasks become heavy, system resources such as the RAM space and CPU time slice etc. are used up. This causes the slow down of the system, and some times, even the traffic jams.

When the chaos structure based queries are made, the CGI programs carry with them the querying parameters for retrieving the chaos structured documents, and transfer the intermediate results and those parameters to the API programs. In this manner, the queries can share the same one CGI program without further instantiation. When a different query task is encountered, only the querying parameter formats are changed in order to deal with the documents in the same chaos structure. Following this unified scheme, the CGI programs can be placed in the system RAM memory. Each time these CGI programs are called and given a different set of parameters, they execute a different query task. The situation that multiple loading of CGI programs and multiple CGI initiations is avoided. The unnecessary memory requirement is reduced to the minimum. Also, when the API background processing is started, the CGI programs have already been released

from the task and in the next query cycle. The CPU time used is consequently reduced to the minimum. Since the API programs and the web server are at the same data processing stage, the system resources they need for their work are much less, and their efficiency is much enhanced. This data processing scheme has satisfied the system requirements, has it also avoided the CGI's shortcomings such as slow operations and system jams caused by slow operations.

4. SYSTEM ACHIEVEMENTS

NOTTnet has been designed as a management model on the World Wide Web. The entire finished system has three major components:

(1) The API modules. These modules are responsible for managing the 32 bit raw data foundation databases which are located on the web server and are full text retrievable.
(2) The CGI programs. Sharing the WWW communication prototypes, these programs are responsible for coordinating the foundation level API modules, and for linking the Internet Explorer (IE) or Netscape browsers.
(3) Generic Internet user terminals loaded with IE/Netscape. These are used as the system I/O for query instructions and results.

Users input query instructions at the IE/Netscape interfaces; the instructions are forwarded to the CGI modules; the CGI modules search and retrieve the queried data from the database foundation which contains the full-text retrievable data; and these query results are transferred back to the IE/Netscape interface to display to the users.

NOTTnet system has completed its full product line. It has a stand-alone PC version on Windows platform, a LAN/WAN version on Windows NT platform, and a Internet version for the WWW environment. This full line of software product can be adopted to a broad range of hardware, operating systems, and network configurations. Various implementation scales are available from single PC, LAN/WAN networks, and all way to the WWW environment.

Its database capability is very powerful. A single database can store up to 60,000 records, 32,000 characters on each record, and a total of 200 million characters. The number of databases is only limited by the storage devices. The system's I/O speed is also impressive: it reaches an input rate of 50 million characters per hour, and a retrieval speed of 0.4 second per 10 million characters.

5. REFERENCES

Diance K.Kovacs, *The Internet Trainer's Guide*
Douglase.Comer, *The Internet Book*
David Sache, Henry Stair, *Hands-On INTERNET*
Harley Hahn, Rick Stout, *The Internet Complete Reference*